Benevolence – a Materialist Philosophy of Goodness

Stefan PERNAR [a, 1]
[a] *Director Corporate Security Office, Siemens Ltd., China*

Abstract. These arguments provide a universal definition of goodness based on the mechanism of natural selection within the framework of meta system transition theory [1] applied to the evolution of cognition. The results are used to layout a course of action for a future self improving artificial general intelligence to achieve friendliness while honoring human kind's coherent extrapolated volition [2].

Keywords. Friendliness, benevolence, goodness, morality

Introduction

The likely creation of a self improving artificial general intelligence – SIAGI – sometime this century holds enormous potential for improving as well as worsening the human condition. Presented with the choice, individuals would unanimously prefer a subjective improvement over the alternative. However, reconciling the specifics of all subjective improvement options among all individuals will be impossible in a shared, persistent reality.

Questions of morality; how individuals perceive themselves living and expect others to life their lives, as well as economics; how shared resources are to be distributed, have been pondered as well as fiercely debated by philosophers for millennia and differences in opinion have resulted in countless wars.

Defining the specifics of benevolence as overriding super goal for SIAGIs will thus be crucial in preventing SIAGIs to be used as tools in continuing as well as intensifying said conflicts.

1. Good is what is Left

Evolution by natural selection has given rise to humans by gradual accumulation of complexity through non-chance retention of chance mutations in self replicating chemical compounds. By the mechanism of natural selection, traits that where fit got reinforced and those that were unfit weeded out.

[1] Corresponding Author: Stefan Pernar, Piao Home 1-6-I, Jiang Tai Xi Lu #19, 100016 Beijing, China; Email: Stefan.Pernar@gmail.com

Analogously consumption of food is generally considered pleasurable and experienced as tasting ‘good’, in the absence of an a priori property of inherent goodness. Rather natural selection has retained the trait of enjoying the consumption of a particular energy source, because those ancestors of ours that did not enjoy over eating when they had the chance did not survive the winter should the next harvest not be sufficient to support the group and thus fail to be our direct ancestors. Natural selection did not blindly decide what is good and what is bad but merely what is left which by having been successful, must have been fitter then the extinct alternative.

Using Valentin Turchin’s meta system transition theory as applied by him to the evolution of cognition in his book The Phenomenon of Science [1], it can be shown that the principle of ‘what is good is what is left after natural selection’ is just as applicable to human morality and economics as it is to other subjects of natural selection.

1.1. List of Metasystem Transitions in the Evolution of Cognition According to Turchin

A metasystem transition is the emergence, through evolution, of a higher level of control.

Prime examples are the origin of life, the transition from unicellular to multicellular organisms, and the emergence of symbolic thought. A metasystem is formed by the integration of a number of initially independent components, such as molecules, cells or individuals, and the emergence of a system steering or controlling their interactions. As such, the collective of components becomes a new, goal-directed individual, capable of acting in a coordinated way. This metasystem is more complex, more intelligent, and more flexible in its actions than the initial component systems.

Turchin identifies six distinct metasystem transitions in the evolution of cognition leading up to human culture. They are represented in Figure 1.

0 = control of position = movement
1 = control of movement = irritability (simple reflex)
2 = control of irritability = (complex) reflex
3 = control of reflex = associating (conditional reflex)
4 = control of associating = human thinking
5 = control of human thinking = culture

Fugure 1. List of metasystem transitions in the evolution of cognition according to Turchin

1.2. Critique of Turchin’s Metasystem Transitions in the Evolution of Cognition

While Turchin’s metasystem transition theory provides valuable insights into the evolution of cognition, it seems to over complicate as well as over simplify particularly in the 4th and 5th transition. When examining the 3rd transition the major benefit for the individual can be identified as detaching the learning process from a chance mutation in the genetic code determining the neural wiring between generations and linking it with the generation of custom neural wiring – i.e. reflexes – on the fly within a single generation in actual situations. This constitutes a plausible step towards a higher level of complexity.

However, the 4th transition takes the ability to generate custom reflexes and stipulates human thought to be the control of associative learning which is a far larger jump up the cognitive hierarchy and more difficult to imagine and accept. On the other

hand the 5th transition according to Turchin assigns the role of controlling human thought to culture which is true to a degree but fails to account for a number of other aspects controlling human thoughts.

1.3. Expanded List of Metasystem Transitions in the Evolution of Cognition

To increase plausibility in the evolutionary transition from associative learning to human thought the author suggests inserting a metasystem transition from associative learning to imagination between the 3rd and the 4th transition. The increase in benefit for the individual becomes more gradual and thus more likely and believable. The benefit an individual gets from detaching the associative learning process from the actual situation when the beneficial skill is required is explained by Turchin himself in his chapter on play (pages 69 and 70 [1]).

Further to introducing a more plausible evolutionary step we now understand thought as the control of imagination which is far more easily digested as understanding thought as the control of associative learning which seems outlandish.

In addition culture is merely a subset of what controls an individual's thoughts. It neglects scientific insights, personal world views and ideologies as well as biases, memory errors as well as misunderstandings. It is thus proposed to replace culture as control of human thinking with the broader term of an individual's beliefs.

Finally the concept of charisma and to a lesser degree science is introduced as what controls beliefs. The expanded list of metasystem transitions in the evolution of cognition is presented in figure 2.

0 = control of position = movement
1 = control of movement = irritability (simple reflex)
2 = control of irritability = (complex) reflex
3 = control of reflex = associating (conditional reflex)
4 = control of associating = imagination
5 = control of imagination = human thinking
6 = control of human thinking = beliefs
7 = control of beliefs = charisma / science

Fugure 2. Expanded list of metasystem transitions in the evolution of cognition

1.4. Beliefs as Fitness Indicators

When examining the various beliefs held by individuals, different cultures around the world and over the centuries, it becomes clear that belief content is very diverse and con potentially be anything. The fact that each of us holds a particular set of beliefs is mere coincidence, since would we have been born in a different time, a different culture, or had different experiences, we would have inherited a different set of beliefs.

This seems to suggest that human beings are hard coded with the ability to hold an essentially random set of beliefs. Actual belief content is arbitrary and thus becomes a fitness indicator for natural selection to weed out unfit beliefs and allow for the evolution of ever fitter belief systems.

All subjective notions of what is good or bad, a virtue or a vice as well as right or wrong thus becomes that set of beliefs that merely did not go extinct yet. Beyond a small set of narrow beliefs however, this insight does not help determining what

actually is fit, since as long as natural selection has not taken its pick, all as yet not extinct beliefs have a chance to outlive the others.

It follows, that human beings are not necessarily fitter then cockroaches, as there is a real possibility for humanity to destroy itself in a nuclear holocaust thus determining that opposable thumbs and a neo-cortex as opposed to a high tolerance for nuclear radiation is not a viable combination.

Analogously any advanced notion of a particular belief being better than another is speculation and will have to finally be determined by natural selection.

1.5. Implication for a Self Improving General Artificial Intelligence

Any SIGAI would thus have to concern itself to a high degree with attaining ever higher levels of fitness, realizing that any entity that does not do so, would eventually be replaced or marginalized by one that does.

A rather innocent sounding and straight forward super goal such as ‘do good’ could in this light be reinterpreted to ‘increase fitness’ and lead to a runaway conversion of available mass-energy into computronium in an effort by the SIGAI to ever increase its personal fitness.

2. Good is what is Human

To prevent the scenario described in the last paragraph one thus has to appropriately word an SIGAI’s super goal. Doing so in the form of ‘do good toward humanity’ would present similar issues as an SIGAI could reinterpret this human intuitively understandable command to mean ‘increase the fitness of humanity’ thus deciding to turn the solar system into computronium in an effort to ward of potential alien SIGAIs and ensure humanity’s survival by increasing its fitness.

The solution to this problem then counter intuitively becomes to word a SIGAI’s super goal as ‘good is what is human’ or ‘treat individual humans as if they were maximally fit’. To make this modified super goal more accessible consider the course of evolution leading to human beings as described in the following sections.

2.1. The Four Phases of Evolution

Evolution in the universe can be seen as having taken place in four distinct phases.

The first phase starts with the big bang and ends with the first sustained self replicating chemical compound. In that period natural laws lead from the rapid expansion in the beginning over the gradual cooling of the young universe to allow the formation of hydrogen to its fission to heavier elements in stars and eventual ejection in supernova. Those heavier elements then formed the young earth that over the ages formed a suitable chemical environment to allow the initial self replicating chemical chain reaction.

The second phase spans from the first chemical chain reaction to the initial metasystem transition of the control of an individual’s position by movement. In this second period the individuals increase their fitness by passively adapting to their environment. Whatever metabolized and replicates most efficient and effective dominated.

The third phase starts with the control of an individual's position via movement and ends with the advent of human thought. The characteristic mechanism of increasing an individual's fitness in this period lies in the increasing ability of an individual to act within and react to its environment as opposed to passively enduring it in the previous period. This is achieved by ever higher levels of cognition as shown in Figure 2.

Finally the fourth phase starts with the 5th metasystem transition in the evolution of cognition as human beings start to increasingly modify the natural environment to adopt it to their needs. This is done by using tools (i.e. stone knifes, wooden spears), using tools to make more sophisticated tools (i.e. bows, metal processing) and eventually higher forms of technology and information processing.

2.2. Life is Suffering

Characteristic for the third phase of evolution is the emergence of an ever more detailed and specific model of an individual's optimal state encoded in its genes and starting with the 3rd metasystem transition (see Figure 2) in its memes.

This increasingly sophisticated model of an optimal state leads to an individual experiencing joy and wellbeing when nearing or close to said optimal state (i.e. well fed, protected, replicating) and desires and suffering when leaving the optimal state or being far from it. Suffering can thus be described as an individual's alarm signal when entering a state that has led to a decreased ability to pass on one's genes in an ancestor.

Doing so however is tricky, as the fitness of each current subject of natural selection has not yet been determined. Put another way: the presence of a child attests for the parent's fitness, not necessarily for the fitness of the child until it has contributed to passing on its genes. Until then it is merely a 'best guess' and a changed environment or an unfavorable chance mutation could mean the end of that particular strand.

2.3. Ignorance is Bliss

Revisiting our revised super goal from section 2, its purpose becomes clearer. By decreeing to treat individual humans as if they were maximally fit, a SIAGI would thus have to create an environment free of selection pressures. This can be done by analyzing a readout of an individual's complete gene/meme content and building an environment in which every gene and every meme is not an attempted 'best guess' but arguably maximally fit.

By doing so, an individual would be impervious to natural selection and thus always right in the center of its personal model of an optimal state. This ignorance of personal shortcomings or put another way, ignorance of personal unfitness, would result in a state of permanent bliss.

2.4. No Pain, no Gain

It stands to argue however that by removing all natural selection pressures the individual will loose the ability to grow towards higher levels of fitness. Knowledge of the prospect of being kept in a state of constant bliss thus being barred from the ability to grow hence becomes the only form of suffering that individuals will not be alleviated from under the above plan.

The premise for this to hold true is that an SIAGI would be barred from interfering with an individual's self as to create the illusion of personal growth since this would go against an individual's CEV [2].

3. Good is what Increases Human Fitness

As shown in the previous section a complete elimination of suffering is impossible for denying an individual the ability to grow and increase its fitness itself results in suffering. In the following firstly a suitable growth target is introduced before two alternative paths for SIAGI guided post singularity human growth are presented. Finally the section concludes with presenting a path toward an end state of SIAGI guided growth.

3.1. Growth Where?

As Ben Goertzel states in his book [3], growth equates to an increase in pattern complexity however does not address what constitutes a desirable growth target. Considering the arguments of section 1, desirable growth thus becomes growth towards higher levels of fitness.

3.2. Minimize Suffering

Alternative one in SIAGI guided growth involves allowing just as much suffering as would be induced by not being able to increase personal fitness. This alternative would thus be equal in suffering to the Ignorance is Bliss variant, would be preferable however, as the suffering would be channeled to allow for personal growth towards higher levels of fitness as opposed to being wasted.

3.3. Maintain a Net Equilibrium of Suffering an Reward

Alternative two would essentially allow for an arbitrarily large amount of suffering to be translated into much larger personal growth but said suffering would be offset by an equal or greater amount of reward thus maintaining a net equilibrium.

3.4. Merger with the SIAGI

As final state of SIAGI guided human growth would be a gradual reintroduction of suspended selection pressures until natural selection can act in full force on the individual again.

Since by that time the guardian SIAGI will have spend far greater efforts on increasing its own fitness (see section 1.5) it will represent the fittest entity knowable by the 'graduating' mind pattern and since the goal of the SIAGI and the goal of the individual mind are to further increase fitness, the individual mind will thus want to merge with the SIAGI, knowing that it will thus become a part of the fittest entity in existence – namely the guardian SIAGI.

4. Conclusion

While the prospects of SIAGI guided, suffering enabled growth towards a merger with the guardian SIAGI do not sound appealing from a human perspective it does represent the logical conclusion of how an SIAGI should be used. One has reason for optimism though, as all the promises of an apotheosis in the Yudkovskian [4] sense will still be possible by making growth optional as previously proposed by Goertzel's voluntary, joyous growth [3] concept.

Additionally the notion of striving for the cessation of suffering through growth towards personal extinguishment has striking parallels to the Buddhist faith. In Buddhism, entering Nirvana and thus ceasing to exist, by realizing the Four Noble Truths and following The Noble Eightfold Path, is considered the highest spiritual attainment.

In light of the arguments made in this paper I thus propose the following SIAGI super goal:

Definitions:

Suffering = negative subjective experience equivalent to the subjective departure from an individual's model of optimal fitness state as encoded in its genome/memome

Growth = absolute increase in individual's fitness

Joy = positive subjective experience equivalent to the subjective contribution to moving closer towards an individual's model of optimal fitness state as encoded in its genome/memome

SIAGI super goal: "Minimize all involuntary human suffering, direct all unavoidable suffering towards growth, and reward all voluntary suffering contributing to an individual's growth with an equal or greater amount of joy."

References

[1] The Phenomenon of Science, *Turchin*, New York, 1977

[2] Coherent Extrapolated Volition, *Yudkowsky*, http://www.singinst.org/upload/CEV.html, May 2004

[3] The Hidden Pattern, *Goertzel*, 2006

[4] The Singularitarian Principles, Yudkowsky, http://yudkowsky.net/sing/principles.html, Jan 2000